Best Practices in Designing Search Engines and Content Management (CM) Systems

Mark Uhart, CKM , CSC

Battle Command Knowledge System

30 October 2008

This is a Workshop

We might actually get something done!

Workshop Outline

- **The Information Search Cunundrum**
- **Net-Centric Data (sharing) Strategy and Guidance**
- **Net-Centric Results**
- **Web Search**
- **Semantics and Entity and Enterprise Search**
- **Entity Extraction**
- **Enterprise Content Management ECM) Principles**
- **Q & A and Sharing Ideas**

- **Public web sites and repositories with & w/o and search engines**
- **DoD web sites, repositories with & w/o and search engines**
- **Organization/unit web site, repositories and search engines**

DoD and Army Net-Centric Strategy

Secure & Trusted

Discoverable

Accessible

Usable

Interoperable

Manageable

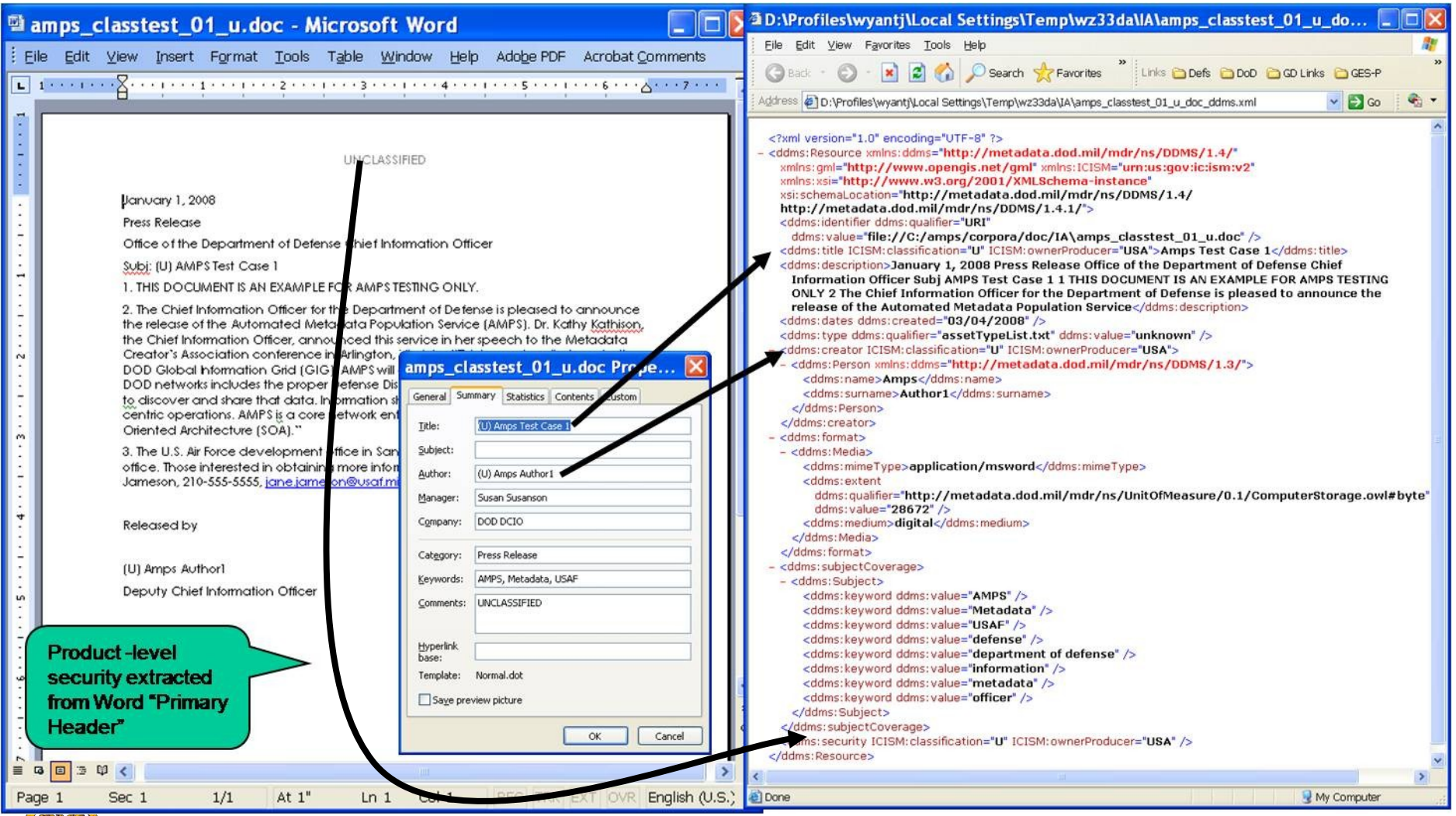
Net-Centric Strategy Guidance &

```
<?xml version="1.0" encoding="UTF-8" ?>
- <ddms:Resource xmlns:ddms="http://metadata.dod.mil/mdr/ns/DDMS/1.4/" xmlns:ICISM="urn:us:gov:ic:ism:v2"
    xmlns:gml="http://www.opengis.net/gml" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
    xsi:schemaLocation="http://metadata.dod.mil/mdr/ns/DDMS/1.4/ http://metadata.dod.mil/mdr/ns/DDMS/1.4.1/">
    <ddms:identifier ddms:qualifier="URI" ddms:value="file://C:/amps/ampsWebServiceProject/corpora\fires_bulletin_p06_may_jun_2008.pdf" />
    <ddms:title ICISM:classification="U"
      ICISM:ownerProducer="USA">C:/amps/ampsWebServiceProject/corpora\fires_bulletin_p06_may_jun_2008.pdf</ddms:title>
    <ddms:description ICISM:classification="U" ICISM:ownerProducer="USA">Fires for the 2007 Surge in Iraq Lethal and Nonlethal Brigadier General Mark
      McDonald Former Deputy Commanding General for Fires and Chief of the Joint Fires Cell, MultiNational Corps, Iraq CPL William
      G</ddms:description>
    <ddms:dates ddms:created="2008-10-21" />
    <ddms:type ddms:qualifier="assetTypeList.txt" ddms:value="unknown" />
  - <ddms:creator ICISM:classification="U" ICISM:ownerProducer="USA">
    - <ddms:Person>
        <ddms:name>General Mark</ddms:name>
        <ddms:surname>McDonald</ddms:surname>
      </ddms:Person>
    </ddms:creator>
  - <ddms:format>
    - <ddms:Media>
        <ddms:mimeType>application/pdf</ddms:mimeType>
        <ddms:extent ddms:qualifier="http://metadata.dod.mil/mdr/ns/UnitOfMeasure/0.1/ComputerStorage.owl#byte"
          ddms:value="1061683" />
        <ddms:medium>digital</ddms:medium>
      </ddms:Media>
    </ddms:format>
  - <ddms:subjectCoverage>
    - <ddms:Subject>
        <ddms:keyword ddms:value="adaptive control system" />
        <ddms:keyword ddms:value="adaptive system" />
        <ddms:keyword ddms:value="agent" />
        <ddms:keyword ddms:value="al qaeda" />
        <ddms:keyword ddms:value="all weather" />
        <ddms:keyword ddms:value="ammunition" />
```

DoD Discovery Metadata Specification

- **The DoD Net-Centric Data Strategy (NCDS) and Directive 8320.2 require data sharing across the DoD, including the creation of new information resources to describe available data:**
- **[POLICY] 4.2. Data assets shall be made visible by creating and associating metadata (“tagging”), including discovery metadata, for each asset. Discovery metadata shall conform to the Department of Defense Discovery Metadata Specification (DDMS). [Department of Defense Directive Number 8320.2 (December 2, 2004), p. 2., directive certified current as of April 23, 2007]**
- **Use of DDMS is required!**
- **http://metadata.dod.mil/mdr/irs/DDMS/#DDMS_info**

Metadata Extraction and ~~Population~~



Metadata Extraction and Population

```
<?xml version="1.0" encoding="UTF-8" ?>
- <ddms:Resource xmlns:ddms="http://metadata.dod.mil/mdr/ns/DDMS/1.4/" xmlns:ICISM="urn:us:gov:ic:ism:v2"
    xmlns:gml="http://www.opengis.net/gml" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
    xsi:schemaLocation="http://metadata.dod.mil/mdr/ns/DDMS/1.4/ http://metadata.dod.mil/mdr/ns/DDMS/1.4.1/">
    <ddms:identifier ddms:qualifier="URI"
      ddms:value="file://C:/amps/ampsWebServiceProject/corpora\1id_1bde_memo_coin_i_poi_version_oef_20070801.doc" />
    <ddms:title ICISM:classification="U" ICISM:ownerProducer="USA">Advance Sheet for COIN I Fundamentals - OEF</ddms:title>
    <ddms:description ICISM:classification="U" ICISM:ownerProducer="USA">AFZN-BA SUBJECT Advance Sheet for COIN I - OEF AFZN-BA 01 Aug 07
      MEMORANDUM FOR RECORD SUBJECT Advance Sheet for COIN I Fundamentals - OEF</ddms:description>
    <ddms:dates ddms:created="2008-10-21" />
    <ddms:type ddms:qualifier="assetTypeList.txt" ddms:value="unknown" />
  - <ddms:creator ICISM:classification="U" ICISM:ownerProducer="USA">
    - <ddms:Person>
        <ddms:name>MAJ Joshua Potter</ddms:name>
      </ddms:Person>
    </ddms:creator>
  - <ddms:format>
    - <ddms:Media>
        <ddms:mimeType>application/msword</ddms:mimeType>
        <ddms:extent ddms:qualifier="http://metadata.dod.mil/mdr/ns/UnitOfMeasure/0.1/ComputerStorage.owl#byte" ddms:value="441856" />
        <ddms:medium>digital</ddms:medium>
      </ddms:Media>
    </ddms:format>
  - <ddms:subjectCoverage>
    - <ddms:Subject>
        <ddms:keyword ddms:value="COIN POI" />
        <ddms:keyword ddms:value="action" />
        <ddms:keyword ddms:value="advance sheet" />
        <ddms:keyword ddms:value="afghanistan" />
        <ddms:keyword ddms:value="afzn-ba" />
        <ddms:keyword ddms:value="analysis model" />
        <ddms:keyword ddms:value="analysis result" />
        <ddms:keyword ddms:value="analysis" />
        <ddms:keyword ddms:value="applicable reference" />
        <ddms:keyword ddms:value="applicable" />
        <ddms:keyword ddms:value="based" />
```

Web Content Discoverability

- **Include visual aids:**
  - Logical and well structured taxonomy that users understand
  - Use of channels to separate content by purpose, type or topic
  - Location of standard features like search tool box and contact links
  - Robust cross-linking to other pages and no dead-end pages
  - Hierarchical and non-hierarchical clues on every page
  - Visited links clearly identified
  - Font sizes accommodate all age groups (not too small)
- **Design in good metadata behind HTML pages:**
  - Use highly–targeted key words (Consider using a Keyword Discovery API)
  - View source (html) code to ensure there is good “title”, keywords” and “description“ information. This is most important for public sites.
  - Include heading tags and alt tags for images.
  - Place any script code into external files.

Web Content Discoverability (cont.)

- **Make sure content is discoverable and usable:**
  - **Author completes document properties:**
    - ✓ **Author:**
    - ✓ **Title:**
    - ✓ **Subject:**
    - ✓ **Comments:**
    - ✓ **Company (Unit):**
    - ✓ **Custom properties for other metadata, e.g. hyperlink, department, mailstop, office symbol project, etc., per SOP**
  - **MS Office files should be backward compatible (.doc vs. .docx).**
  - **PDF files must be text-readable.**
  - **MS Office Restriction Permissions = Unrestricted access**
  - **PDF Security Method = No Security**

Document Properties

Would you live in a home without an address?

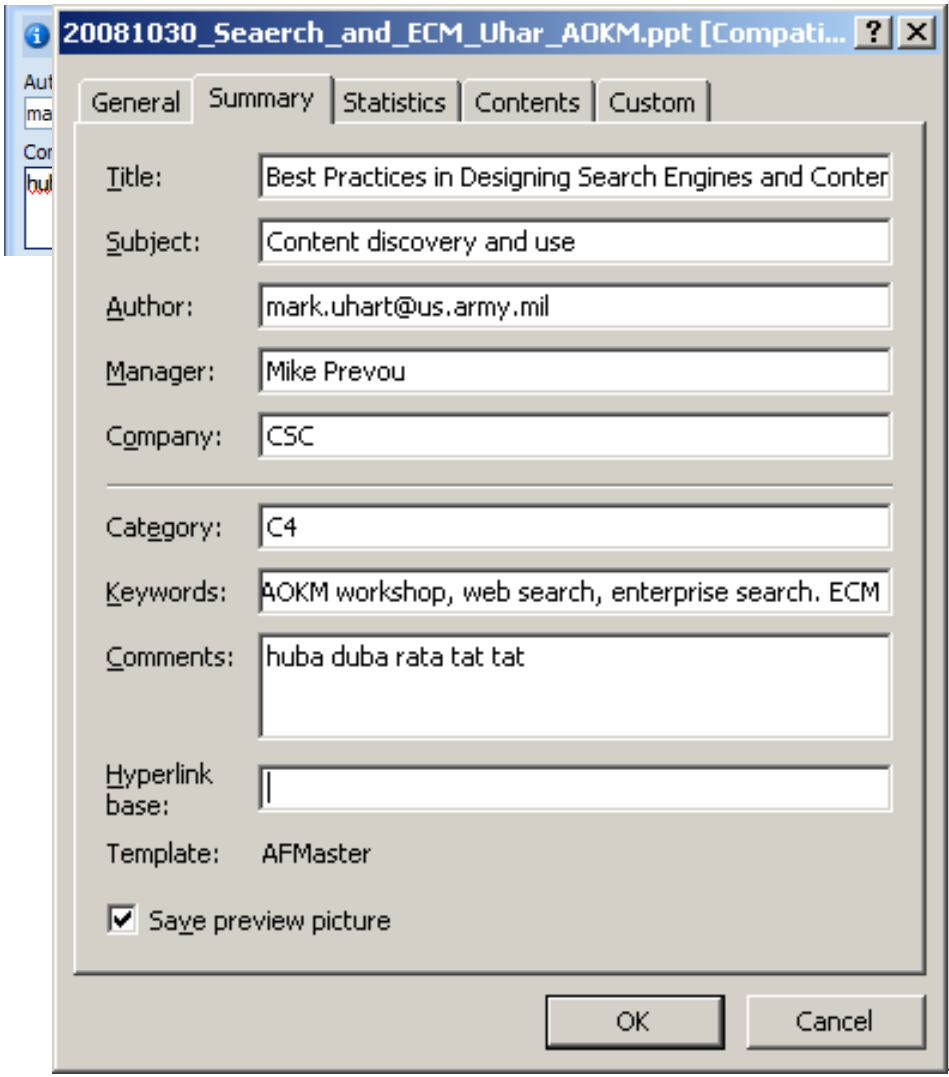
Would you have a pet without a name?

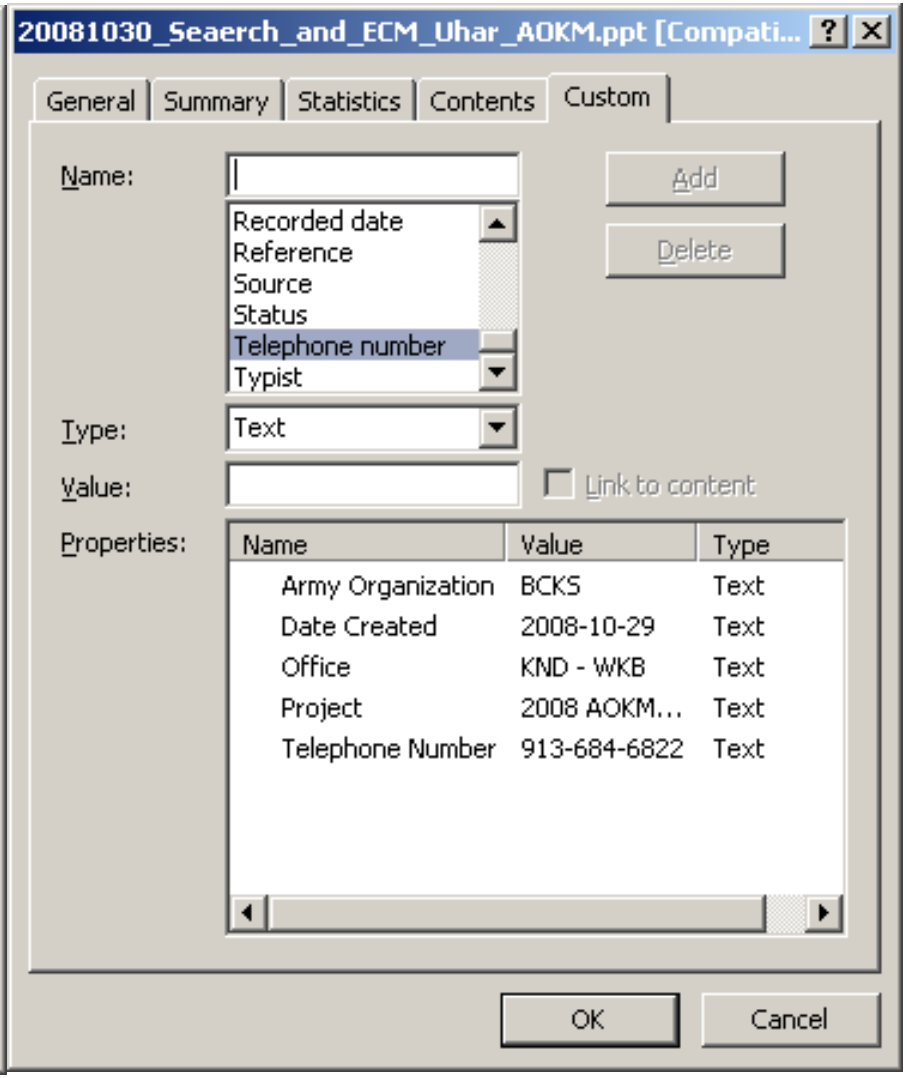
Would you drive a car without a license plate number?

Would you draw Social Security without a SSN?

Then why would you create a document or file without a means to find it?

Document Properties

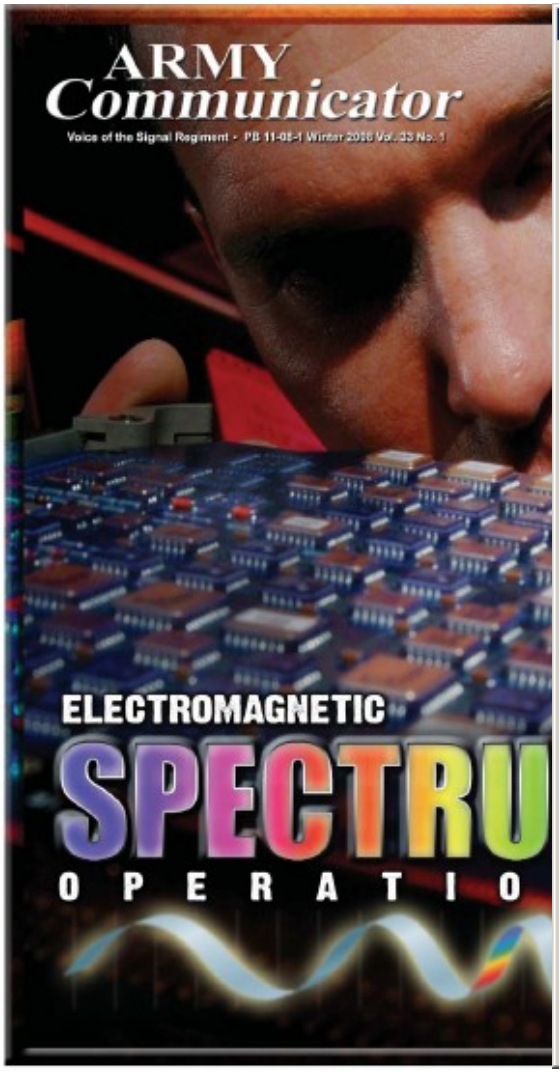


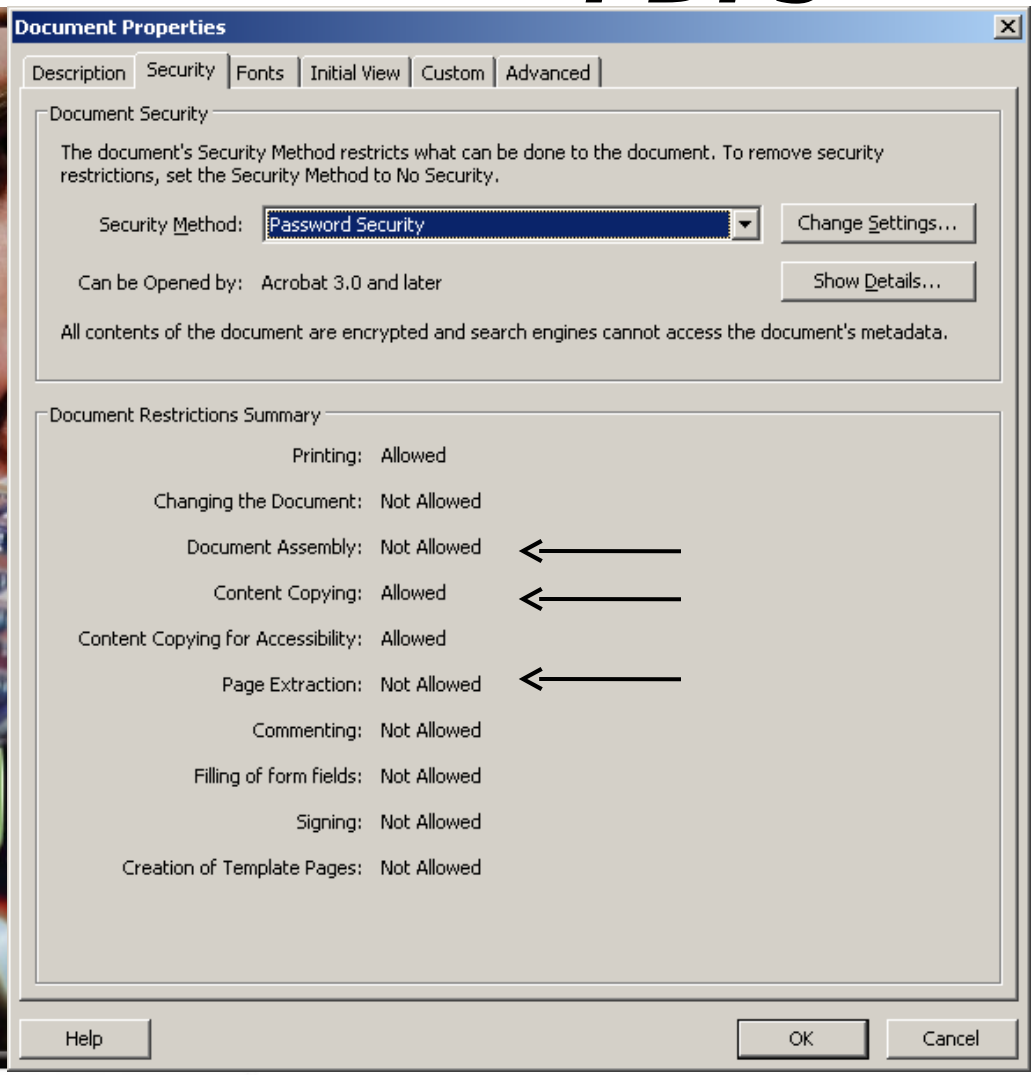


Summary properties/metadata

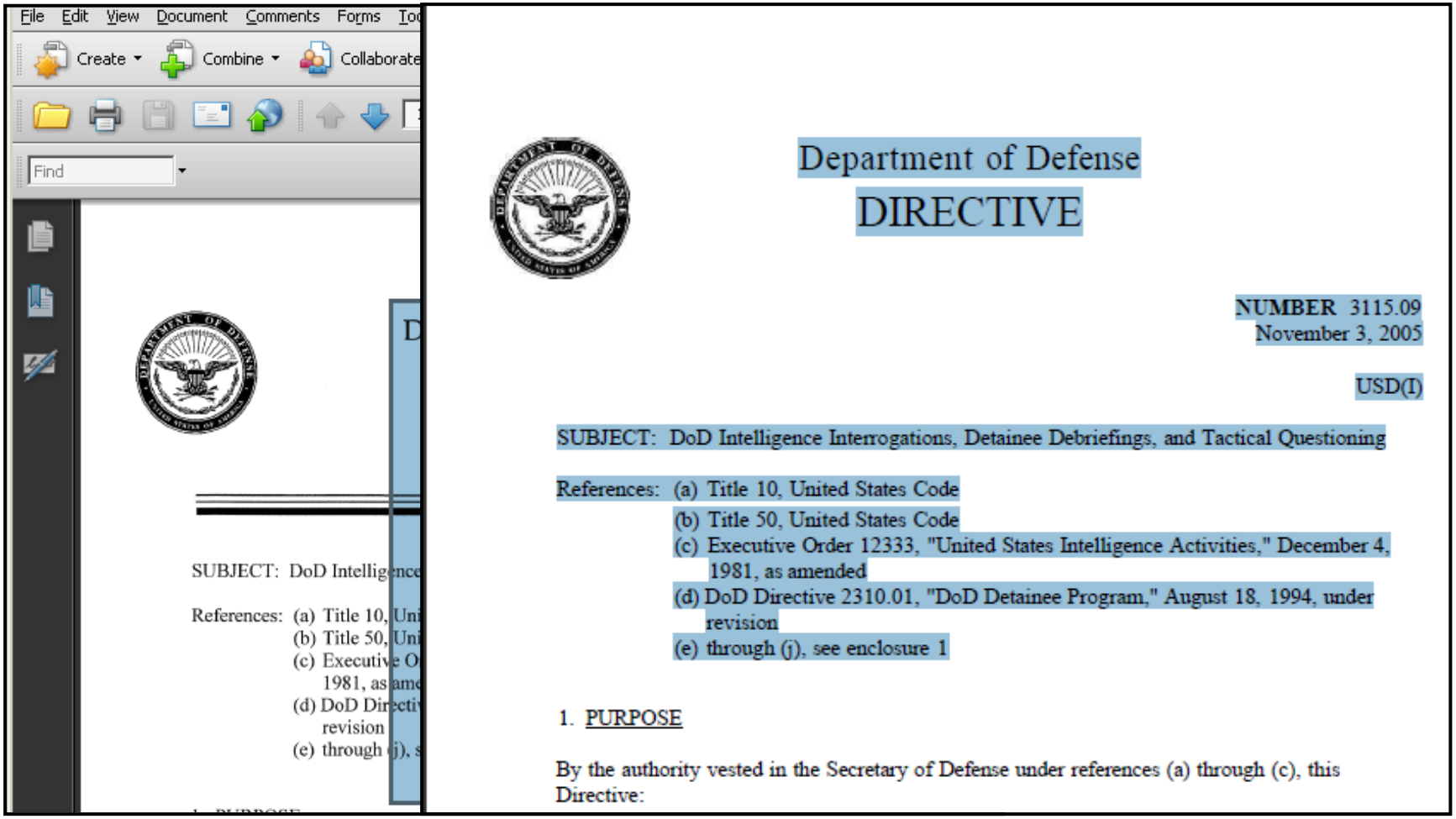
Custom properties/metadata

Security Restrictions on ~~PDFs~~



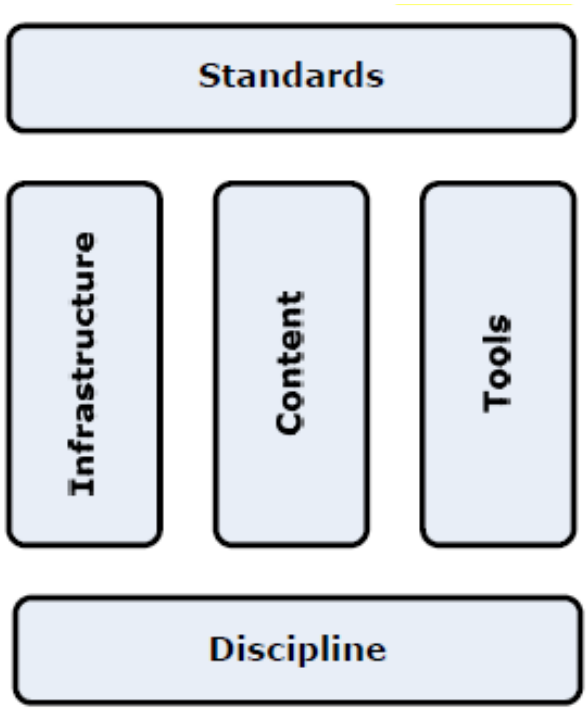

Document Properties
Description
Security
Fonts
Initial View
Custom
Advanced
Document Security
The document's Security Method restricts what can be done to the document. To remove security restrictions, set the Security Method to No Security.
Security Method: Password Security
Change Settings...
Can be Opened by: Acrobat 3.0 and later
Show Details...
All contents of the document are encrypted and search engines cannot access the document's metadata.
Document Restrictions Summary
Printing: Allowed
Changing the Document: Not Allowed
Document Assembly: Not Allowed
Content Copying: Allowed
Content Copying for Accessibility: Allowed
Page Extraction: Not Allowed
Commenting: Not Allowed
Filling of form fields: Not Allowed
Signing: Not Allowed
Creation of Template Pages: Not Allowed
Help
OK
Cancel

You can't read a picture.

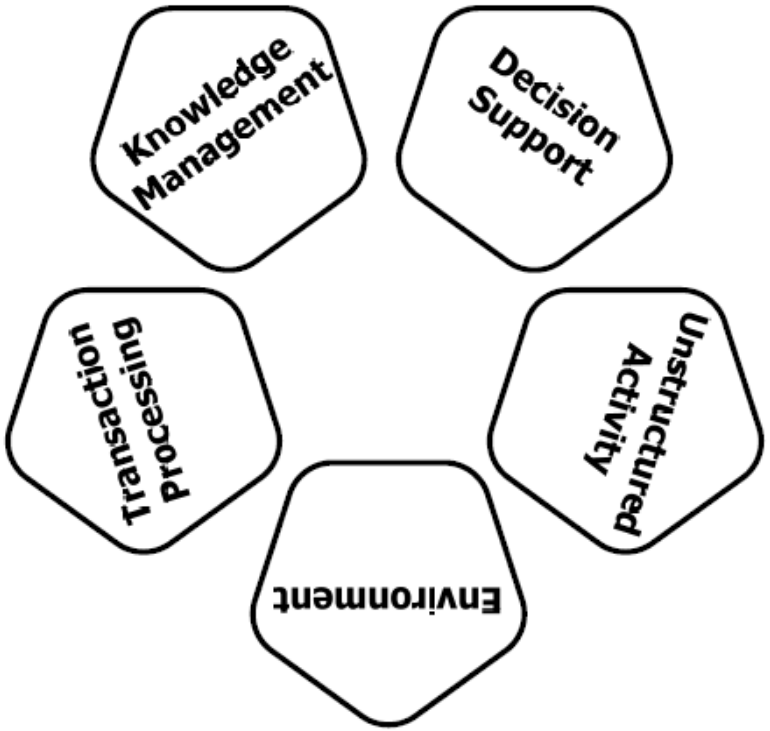

File Edit View Document Comments Forms
Create
Combine
Collaborate
Find
Department of Defense
DIRECTIVE
NUMBER 3115.09
November 3, 2005
USD(I)
SUBJECT: DoD Intelligence Interrogations, Detainee Debriefings, and Tactical Questioning
References: (a) Title 10, United States Code
(b) Title 50, United States Code
(c) Executive Order 12333, "United States Intelligence Activities," December 4, 1981, as amended
(d) DoD Directive 2310.01, "DoD Detainee Program," August 18, 1994, under revision
(e) through (j), see enclosure 1
1. PURPOSE
By the authority vested in the Secretary of Defense under references (a) through (c), this Directive:

But the creators of documents think the software can?

Understanding Semantics



Semantic Framework



Semantics Applied Differently

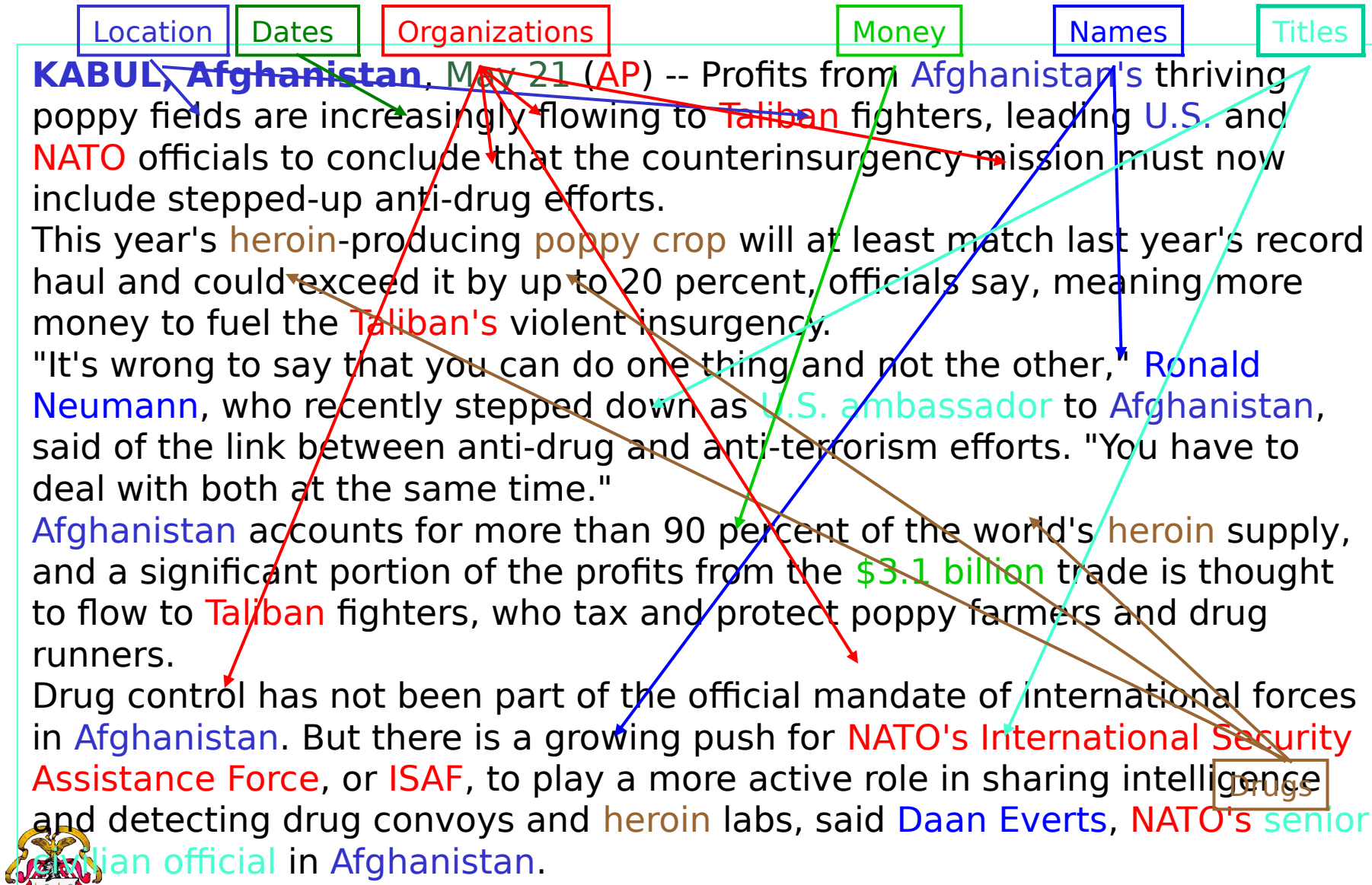
Extracts from the CIO's Guide to Semantics, Version 2, by Sematic Arts at: http://www.wilshireconferences.com/wilshireconf_cfmfiles/stc06/PDF_file_request1.cfm

Search Engine Design Recommendations

- Apply a DDMS-compliant schema, mark-up files in XML and use discovery metadata.
- Use the “and” as the default operator. For example, when searching for “civil information management,” search for:
  1. “civil” + “information” + “management” first, as linked words;
  2. “civil” + “information”, “civil” + “management”, “information” + “management”, and “civil” + “management” second, as linked words;
  3. “civil,” “information,” and “management “ not linked but on the same page or in the same document; and
  4. “civil” or “information” or “management” anywhere in the document.
- Review and apply stop words – a, an, and, but, can, do, for, he, etc.
- Apply word stemming – command > commander, commanded, commanding
- Design for semantic discovery by applying:
  - the English Dictionary so words are suggested when keywords are incorrectly spelled;
  - An Army dictionary of terms such as the ABCA;
  - A COI controlled vocabulary (dictionary and thesaurus).
- Provide filtering by metadata, e.g. author, title, file type, subject/category or date created.
- Always open search results in a new browser.
- Use web logs to collect user behavior data and build in metrics collection capability.

Design it well and they will come.

Entity Extraction


Location
Dates
Organizations
Money
Names
Titles
KABUL, Afghanistan, May 21 (AP) -- Profits from Afghanistan's thriving poppy fields are increasingly flowing to Taliban fighters, leading U.S. and NATO officials to conclude that the counterinsurgency mission must now include stepped-up anti-drug efforts.
This year's heroin-producing poppy crop will at least match last year's record haul and could exceed it by up to 20 percent, officials say, meaning more money to fuel the Taliban's violent insurgency.
"It's wrong to say that you can do one thing and not the other," Ronald Neumann, who recently stepped down as U.S. ambassador to Afghanistan, said of the link between anti-drug and anti-terrorism efforts. "You have to deal with both at the same time."
Afghanistan accounts for more than 90 percent of the world's heroin supply, and a significant portion of the profits from the $3.1 billion trade is thought to flow to Taliban fighters, who tax and protect poppy farmers and drug runners.
Drug control has not been part of the official mandate of international forces in Afghanistan. But there is a growing push for NATO's International Security Assistance Force, or ISAF, to play a more active role in sharing intelligence and detecting drug convoys and heroin labs, said Daan Everts, NATO's senior civilian official in Afghanistan.
Drugs

ECM Overview

Enterprise Content Management

What is ECM?

The strategies, methods and tools used to **capture**, **manage**, **store**, **preserve** and **deliver** content and documents related to organizational processes. [1]

A set of technologies used to capture, store, preserve and deliver content and documents related to organizational processes. ECM tools and strategies allow the management of an organization's unstructured information, wherever that information exists. [2] NOT A GOOD DEFINITION – EXCLUDES PEOPLE AND PROCESSES.

ECM is not:

- technology-driven or based on a single technology
- a panacea for managing all explicit content
- easy

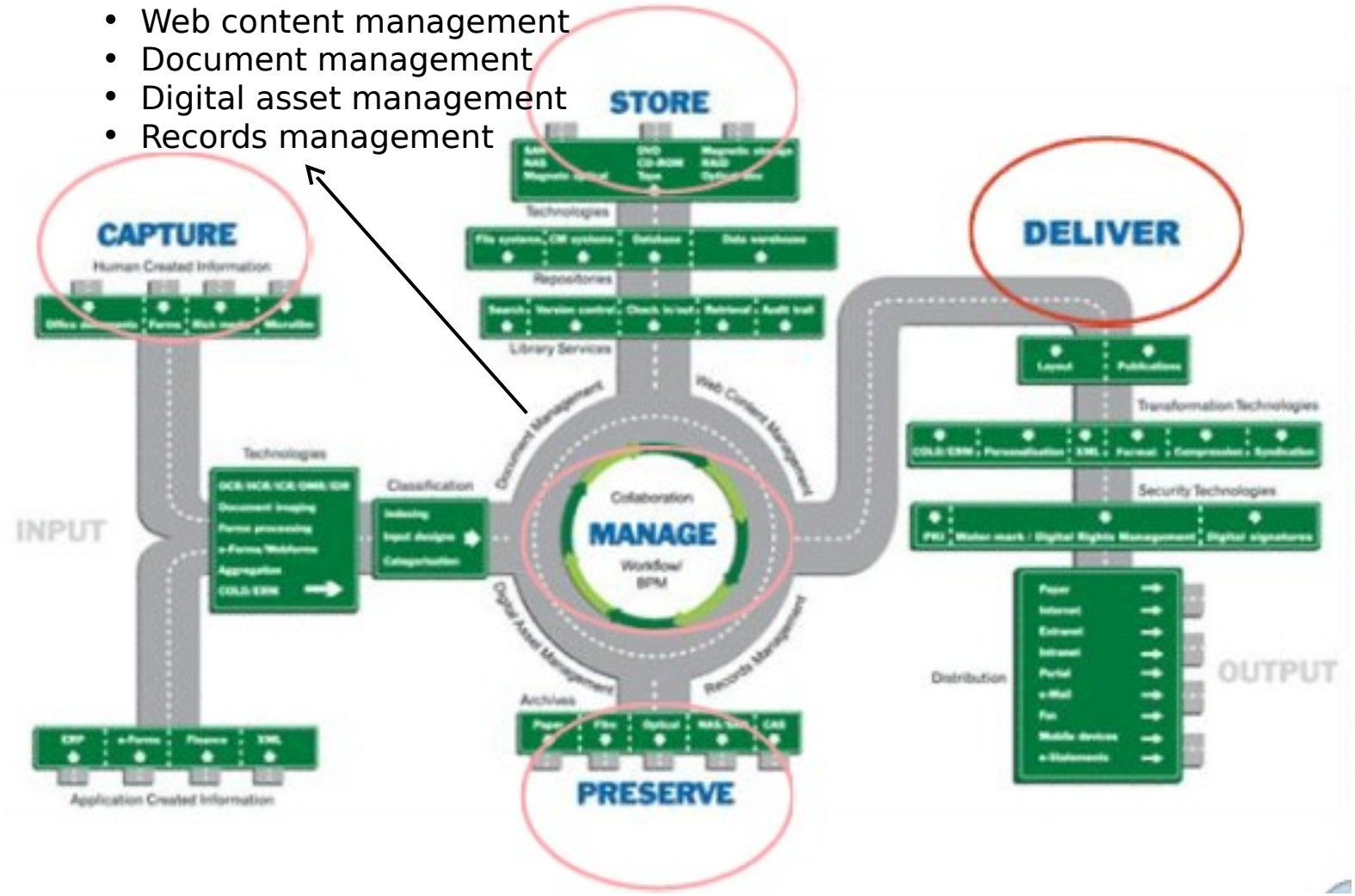
ECM is:

- about people and organizational and functional area processes and workflow;
- about integrating structured and unstructured data/content from many sources;
- complicated and requires a great deal of governance, planning and structure.

1 - Definition from AIIM **- *http://www.aiim.org***

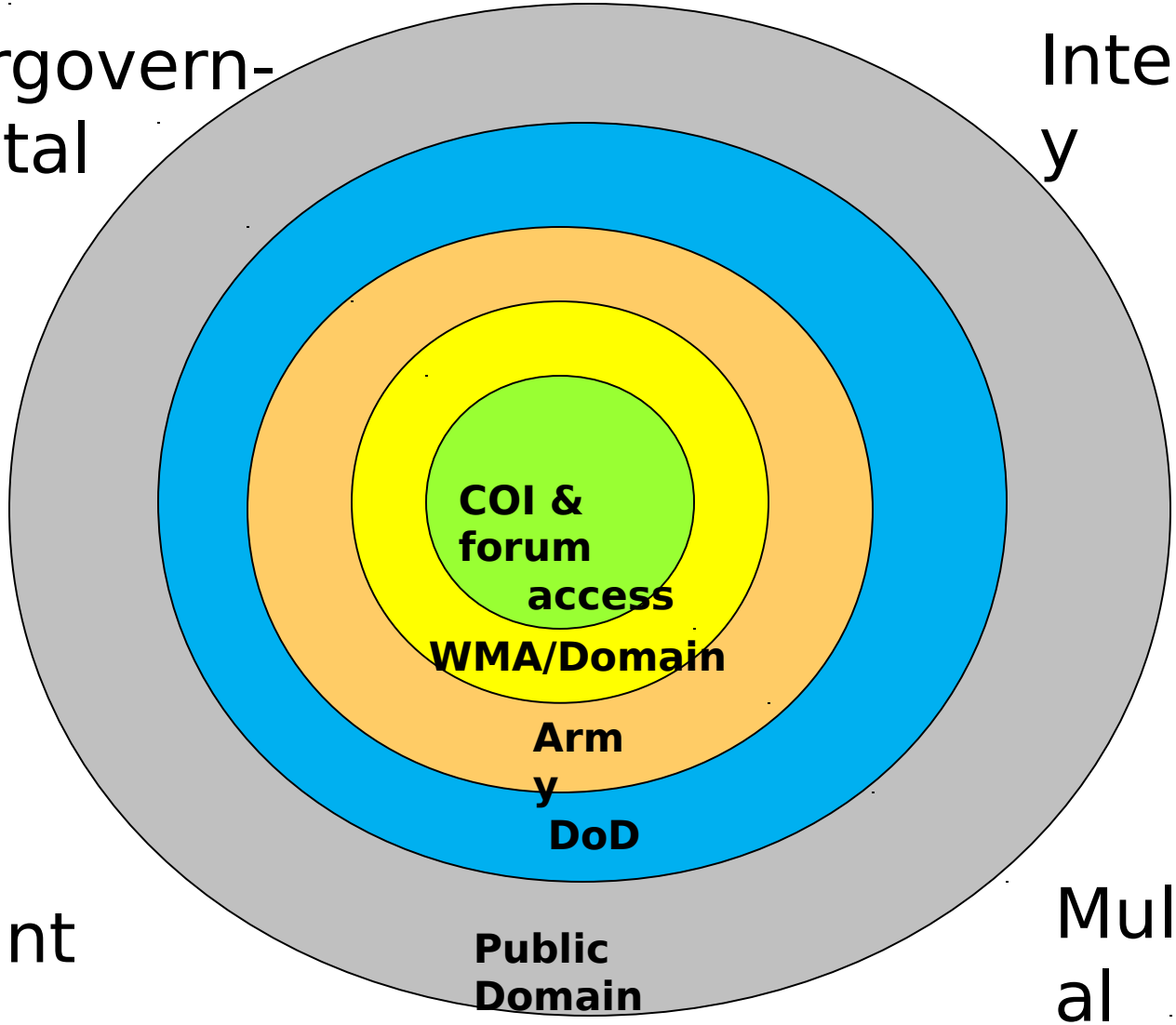
2 – Definition from Wikipedia

- Web content management
- Document management
- Digital asset management
- Records management



Extract from AIIM ECM Practitioner Certificate Program

Access Rights



ECM Model Planning Considerations

1. Governance, authority and policies (commanders, staff, NCOs, enlisted, groups and teams)
2. Legislation/law, regulation and standards - FOIA, Privacy, Sect 508, HIPAA, Sarbanes-Oxley
3. Classification – structured and unstructured data; sharability (domain, COI, COP), ontology and taxonomy, record/non-record; static, dynamic or mixed content, genre and file types, single or multiple collections, file management and content indexing
4. Controls and Security:
   - Administration - user and admin privileges, access by roles/rights/affiliation)
   - Ownership and integrity - authenticated source, version control, encryption
   - Access rights – JIIM and NGO considerations, classification, dissemination controls, copyright controls, trust and privacy
   - Security – data protection and back-up, PKI and electronic signatures, security markings, OPSEC
   - Interoperability – JIIM, NGOs, military alliances
5. Strategy, processes and workflow for capturing, managing, storing, preserving and delivering information, e.g. parsing, rendering, discovering, retrieving, repurposing
6. Interfaces and linkage – JIIM and NGOs, collaborative systems, web CM, data asset repositories and databases, legacy/non net-centric systems, workflow tools and applications
7. Standards – W3C (OWL, HTML, XML), schemas and metadata (DDMS,

AIIM ECM Architecture

Extract from AIIM International and Doculabs ECM 101 Poster

What's Next

- Q & A
- Others share their search and ECM experience

Workshop Outline

- **The Information Search Cunundrum**
- **Net-Centric Data (sharing) Strategy and Guidance**
- **Net-Centric Results**
- **Web Search**
- **Semantics and Entity and Enterprise Search**
- **Entity Extraction**
- **Enterprise Content Management ECM) Principles**
- **Q & A and Sharing Ideas**

Let's not forget why we are here.